ANALISI IN COMPONENTI PRINCIPALI (ACP)

Analisi multivariata

a.a. 2011/2012

Carla Rampichini

rampichini@ds.unifi.it

Origini dell' ACP

- Pearson (1901)

- Hotelling (1933)

Una delle tecniche multivariate più utilizzate.

Dove studiare ACP

- Zani S. e Cerioli A. (2007) *Analisi dei dati e data mining per le decisioni aziendali*, Milano: Giuffrè, cap. 6
- dispense on-line Marchetti
- Johnson R.A. e Wichern D.W. (2007). *Applied Multivariate Statistical Analysis*. Sixth Edition. Pearson Education International. Cap. 8
- Per approfondire: Jolliffe, I. T. (2002). *Principal component analysis* , Springer.

Quando è utile l'ACP?

- Analisi di regressione in caso di collinearità;
- problemi di classificazione per gruppi ben separati;
- **riduzione delle dimensioni;**
- identificazione degli outliers.

Cosa fa l'ACP?

Matrice dei dati *Originale* $n * p$

Nuova matrice dei dati $n * q$, con $q < p$

altri risultati utili

Come sono i *nuovi* dati?

Ogni nuova variabile è combinazione lineare delle variabili originali

Le nuove variabili sono scelte in modo da catturare al massimo la variabilità presente nei dati originali

le nuove variabili sono incorrelate!

Le nuove variabili sono dette "scores" o "componenti principali"

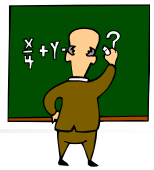
Strumenti per la *comprensione* della ACP

Matrice di Varianza/Covarianza

Autovalori (Eigenvalues) e Autovettori (Eigenvectors)

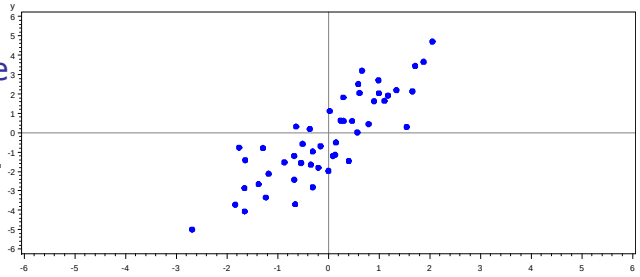
Operazioni elementari su matrici

Componenti principali



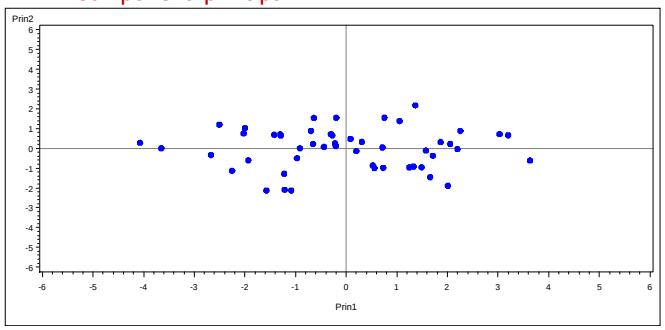
- Algebricamente le componenti principali (scores) sono combinazioni lineari delle variabili $X_1,\ldots,X_p$
- Geometricamente le CP rappresentano un nuovo sistema di coordinate ottenuto ruotando gli assi originali che hanno $X_1,\ldots,X_p$ come coordinate
- I nuovi assi rappresentano le direzioni principali (autovettori o eigenvectors), cioè le direzioni di variabilità minima
- Le CP forniscono una descrizione più parsimoniosa della struttura di covarianza

Un primo esempio: dai dati originali alle CP



Dati generati da Normale bivariata rho=0.8

Componenti principali



C'è più variabilità in direzione di PRIN1 che in X e Y, ma poca variabilità lungo PRIN2

Dai dati originali alle CP

- In generale, da un insieme di p>2 variabili correlate, le prime CP tengono conto della maggior parte della variabilità nelle variabili originali
- Viceversa, le ultime CP identificano direzioni lungo le quali c'è poca variabilità, cioè relazioni lineari quasi costanti tra le variabili originali

Vedi acp_dettagli.pdf, slides 1-9

CP della popolazione (1)

- Sia $\mathbf{x}'=[X_1,\ldots,X_p]$ un v.c con matrice di covarianza $\boldsymbol{\Sigma}$ e autovalori $\lambda_1 \geq \lambda_2 \ldots \geq \lambda_p$
- Consideriamo le p combinazioni lineari $Y_k=\mathbf{a}_k'\mathbf{x}=a_{k1}X_1+\ldots+a_{kp}X_p,\ k=1,\ldots,p$

- $\mathrm{Var}(Y_k)=\mathbf{a}_k'\,\boldsymbol{\Sigma}\mathbf{a}_k,\ k=1,\ldots,p$
- $\mathrm{Covar}(Y_l,Y_k)=\mathbf{a}_l'\,\boldsymbol{\Sigma}\mathbf{a}_k,\ l,k=1,\ldots,p$

Ipotesi di base

- Le CP dipendono solo dalla matrice di covarianza $\boldsymbol{\Sigma}$ (o dalla matrice di correlazione)
- Non occorre fare ipotesi sulla distribuzione congiunta delle v.c.
- Se però $\mathbf{x}'=[X_1,\ldots,X_p] \sim MN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ le CP hanno un'utile interpretazione in termini di ellissoidi di iso-densità

La prima CP

- La prima CP è la combinazione lineare $Y_1=\mathbf{a}_1'\mathbf{x}$ tale che

$$\text{Var}(Y_1)=\mathbf{a}_1'\mathbf{\Sigma a}_1=\max$$

  - Questa quantità può essere resa grande a piacere moltiplicando $\mathbf{a}_1$ per una costante
  - → si impone il vincolo di lunghezza unitaria:

$$\mathbf{a}_1'\,\mathbf{a}_1=1$$

Come ottenere $V(Y_1)$ massima?

cerchiamo i coefficienti $\mathbf{a}_1$:

$$\max V(Y_1)=\mathbf{a}_1'\mathbf{\Sigma a}_1,\ \text{s.t. } \mathbf{a}_1'\mathbf{a}_1=1$$

Soluzione

Moltiplicatori di Lagrange (Morrison, 1990)

radici caratteristiche dell'equazione:

$$|\mathbf{\Sigma}-\lambda\mathbf{I}|=\mathbf{0}$$

Autovettori: $\mathbf{\Sigma e}_j=\lambda_j\mathbf{e}_j$

Dimostrazione alternativa: (2-51) p.80 JW

La prima componente principale

combinazione lineare delle X originali:

$$Y_1=e_{11}X_1+\ e_{12}X_2+\ +\ldots+\ e_{1p}X_p=\mathbf{e}'_1\mathbf{x}$$

Notare che:

$$Var(Y_1)=\mathbf{e}_1'\mathbf{\Sigma e}_1=\mathbf{e}_1'\mathbf{E\Lambda E}'\mathbf{e}_1=\lambda_1$$

Ricordiamo SVD: $\mathbf{\Sigma=E\Lambda E}'$
$\mathbf{E}=[\mathbf{e}_1,\ldots,\mathbf{e}_p]$, $\mathbf{\Lambda}=\text{diag}(\lambda_k)$

...le altre CP

$\mathbf{a}_k$: $\max V(Y_k)=\mathbf{a}_k'\mathbf{\Sigma a}_k=\max$

s.t. $\mathbf{a}_k'\mathbf{a}_k=1$ e $\text{Cov}(\mathbf{a}_l,\mathbf{a}_k)=0$

- Si ottengono le combinazioni lineari

$$Y_k=e_{k1}X_1+\ e_{k2}X_2+\ \ldots+\ e_{kp}X_p=\mathbf{e}'_k\mathbf{x}$$

- Tra loro incorrelate e con

$$Var(Y_k)=\mathbf{e}_k'\mathbf{\Sigma e}_k=\mathbf{e}_k'\mathbf{E\Lambda E}'\mathbf{e}_k=\lambda_k$$

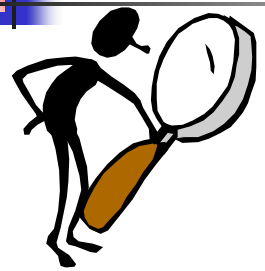
CP della popolazione (2)

- Le CP della popolazione sono le combinazioni lineari $Y_1,\ldots Y_p$

$$Y_k=\mathbf{a}_k'\mathbf{x}=a_{k1}X_1+\ldots+a_{kp}X_p$$

  - incorrelate $\text{Covar}(Y_l,Y_k)=\mathbf{a}_l'\,\Sigma\mathbf{a}_k=0\ l\neq k$
  - e con $\text{Var}(Y_k)=\mathbf{a}_k'\Sigma\mathbf{a}_k$, $k=1,\ldots,p$ più grande possibile

Strumenti per *utilizzare* la ACP



SAS

proc princomp;

proc factor;

STATA

pca

factor

esempi http://www.ats.ucla.edu/stat/sas/examples/cama4/default.htm

Esempio
Vole Data

- Dati morfometrici sulle specie Microtus
- otto varaibili misurate su teschi ritrovati in Europa centrale
- le prime tre variabili sono la lunghezza dei molari in alto a sinistra 1,2 e 3
- dati analizzati da Flury e Salvioni nel 1996

SAS Program

```
data vole;input id y1-y8;cards;
     1   2078   1649   1708   3868   5463   2355    805    475
     1   1929   1551   1550   3825   4741   2305    760    450
     1   1888   1613   1674   4440   4807   2388    775    460
     1   2020   1670   1829   3800   4974   2370    766    460
...

     1   2023   1612   1781   4148   4940   2340    796    455
     1   1885   1549   1628   3718   5286   2300    810    455
     1   1945   1580   1739   3801   5567   2370    800    475
     1   2186   1847   1896   4160   5587   2470    845    500
     1   2110   1631   1703   3856   4773   2350    850    465
;
data vole;set vole (keep = y1-y3);
          rename y1 = lnum1
                 y2 = lnum2
                 y3 = lnum3 ;

proc princomp covariance out=scores;
var lnum1-lnum3;
title 'Default Output of Proc Princomp';


proc print data=scores;
title 'Principal Components (Scores) for Vole Data';
```

vole.sas

SAS Output

Simple Statistics

	LNUM1	LNUM2	LNUM3
Mean	2054.534884	1636.465116	1819.930233
StD	130.962981	101.147782	152.502052

Covariance Matrix

Covarianza

	LNUM1	LNUM2	LNUM3
LNUM1	17151.30233	9741.07863	12121.51440
LNUM2	9741.07863	10230.87375	11037.05703
LNUM3	12121.51440	11037.05703	23256.87597

Total Variance = 50639.052049

quante componenti?

Eigenvalues of the Covariance Matrix

	Eigenvalue	Difference	Proportion	Cumulative
PRIN1	39970.0	32137.4	0.789311	0.78931
PRIN2	7832.5	4996.0	0.154674	0.94399
PRIN3	2836.5	.	0.056015	1.00000

SAS Output

Direzioni

Eigenvectors

	PRIN1	PRIN2	PRIN3
LNUM1	0.560728	0.726846	-.396584
LNUM2	0.443228	0.141076	0.885238
LNUM3	0.699380	-.672155	-.243053

Nuovi Dati

Principal Components (Scores) for Vole Data

OBS	LNUM1	LNUM2	LNUM3	PRIN1	PRIN2	PRIN3
1	2078	1649	1708	-59.568	94.058	28.995
2	1929	1551	1550	-297.055	78.133	39.736
3	1888	1613	1674	-205.842	-26.268	80.742
4	2020	1670	1829	1.842	-26.467	41.178
...						
40	1885	1549	1628	-268.062	-6.558	36.457
41	1945	1580	1739	-143.047	-33.183	13.125
42	2186	1847	1896	220.233	74.126	115.748
43	2110	1631	1703	-53.100	118.139	1.586

osservare .561*(2078-2054.53)+.443*(1649-1636.47)+.699*(1708-1819.93) = -59.5

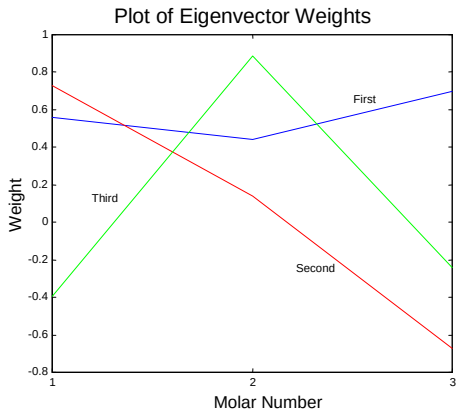
SAS Output

Le nuove variabili sono incorrelate

Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0 / N = 43

	PRIN1	PRIN2	PRIN3
PRIN1	1.00000 0.0	0.00000 1.0000	0.00000 1.0000
PRIN2	0.00000 1.0000	1.00000 0.0	0.00000 1.0000
PRIN3	0.00000 1.0000	0.00000 1.0000	1.00000 0.0

Plot coefficienti degli autovettori

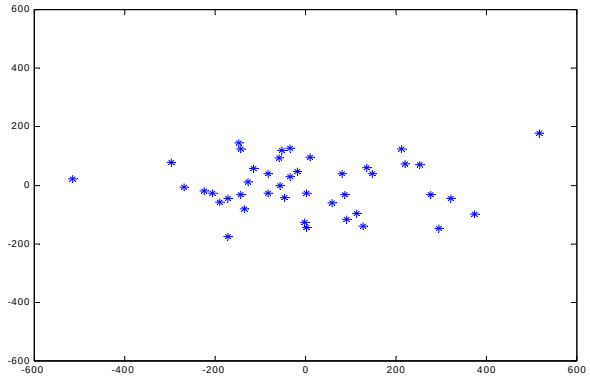


Interpretazione

- La prima componente principale è ottenuta come combinazione lineare con coefficienti tutti positivi delle variabili di dimensione. Possiamo quindi dire che la prima CP misura la 'dimensione' del cranio.

- La seconda nuova variabile si ottiene contrastando la lunghezza del primo molare con quella del terzo. Possiamo quindi dire che la seconda CP è una misura della 'forma' del cranio.

Plot delle osservazioni in base alle prime due CP



PROPRIETÀ

$Y_k = e_{k1}X_1 + e_{k2}X_2 + \ldots + e_{kp}X_p = \mathbf{e}'_k\mathbf{x}$

(1) $\text{Var}(Y_k) = \lambda_k$

(2) $\text{Corr}(Y_k, X_j) = [e_j\sqrt{\lambda_1}]/\sqrt{\sigma_{jj}}$ $\quad \sigma_{jj} = \text{Var}(X_j)$

(3) $\Sigma \lambda_j = \Sigma \sigma_{jj}$, $j=1, \ldots, p$

(4) λ e $\mathbf{e}$ sono gli autovalori e autovettori della matrice di covarianza: $\Sigma\mathbf{e}_k = \lambda\mathbf{e}_k$

(5) autovettori ortogonali: $\mathbf{E}'\mathbf{E} = \mathbf{I}$, $\mathbf{e}_j'\mathbf{e}_k = 0$

(6) Se si utilizza la matrice di correlazione

➔ ACP su variabili standardizzate

Componenti principali (popolazione)

- Dato il vettore casuale $\mathbf{x}' = [X_1, \ldots, X_p]$ con vettore delle medie $\boldsymbol{\mu}$ e matrice di covarianza $\boldsymbol{\Sigma}$, con:
  - autovalori $\lambda_1, \ldots, \lambda_p \geq 0$, $\boldsymbol{\Lambda} = \text{diag}(\lambda_k)$
  - autovettori associati $\mathbf{e}_1, \ldots, \mathbf{e}_p$; $\mathbf{E} = [\mathbf{e}_1, \ldots, \mathbf{e}_p]$
  - poiché $\boldsymbol{\Sigma}\mathbf{E} = \boldsymbol{\Lambda}\mathbf{E}$ per costruzione e $\mathbf{E}$ è ortonormale
    - $\mathbf{E}'\boldsymbol{\Sigma}\mathbf{E} = \boldsymbol{\Lambda}$
    - scomposizione spettrale: $\boldsymbol{\Sigma} = \mathbf{E}\boldsymbol{\Lambda}\mathbf{E}'$
- Le componenti principali di $\mathbf{x}$ sono gli elementi del vettore casuale $\mathbf{y} = \mathbf{E}'(\mathbf{x} - \boldsymbol{\mu})$
- La *k*-ma CP è $y_k = \mathbf{e}'_k(\mathbf{x} - \boldsymbol{\mu})$ con $E(\mathbf{y}) = 0$ e $V(\mathbf{y}) = \lambda_k$

Proprietà 1

- La varianza totale delle CP è uguale alla varianza totale delle variabili originali X

$$\sum_{k=1}^{p}\sigma_{kk} = tr(\boldsymbol{\Sigma}) = tr(\mathbf{E}\boldsymbol{\Lambda}\mathbf{E}') = tr(\boldsymbol{\Lambda}\mathbf{E}'\mathbf{E}) = tr(\boldsymbol{\Lambda}) = \sum_{k=1}^{p}\lambda_k$$

- La proporzione di varianza totale spiegata è pari a
  - dalla k-ma CP $\lambda_j / \left(\sum_{j=1}^{p}\lambda_j\right)$
  - dalle prime k CP $\dfrac{\sum_{j=1}^{k}\lambda_j}{\sum_{j=1}^{p}\lambda_j}$

Proprietà 2

- Per ogni intero q, $1 \le q \le p$, si consideri la trasformazione lineare ortonormale

$$\mathbf{y} = \mathbf{Bx}$$

  - con **y** vettore di dimensione q e B matrice $q \times p$
  - $\Sigma\mathbf{y} = \mathbf{B\Sigma B}$, matrice di covarianza di **y**.

la varianza totale di y, cioè tr(Σ**y**), è massimizzata per $\mathbf{B}=\mathbf{E}_q$, dove $\mathbf{E}_q$ sono le prime q colonne di **E**.

Proprietà 3

- Si consideri nuovamente la trasformazione ortonormale **y** = **Bx**
  - con **x**, **B**, **E** e Σy definite come in precedenza.

Allora la varianza totale di **y**, cioè tr(Σ**y**), è minimizzata considerando

$$\mathbf{B} = \mathbf{E}_{*q}$$

- dove $\mathbf{E}_{*q}$ sono le ultime q colonne di **E**.

Implicazioni statistiche proprietà 3

- Le ultime CP hanno varianze che sono le più piccole possibile.
- Possono quindi essere utilizzate:
  - per trovare relazioni lineari quasi costanti tra le variabili
  - nella regressione
  - per selezionare sottoinsiemi di variabili
  - per trovare outlier.

Proprietà 4

Autovettori $\mathbf{e}_j=[e_{j1}, \ldots, e_{jp}]'$

- gli e_{jk} sono proporzionali alla correlazione tra X_k e Y_j ➔ il loro ordine di grandezza è una misura del contributo della *k*-ma variabile alla *j*-ma CP

$$\rho_{y_j x_k} = \frac{e_{jk}\sqrt{\lambda_j}}{\sqrt{\sigma_{kk}}}$$

- Si ottiene ponendo

$$X_k = \mathbf{a}_k'\mathbf{x}, \quad \mathbf{a}_k' = [0,\ldots 0,1,0,\ldots,0]$$

$$Cov(X_k, Y_j) = E(\mathbf{a}_k'\mathbf{x}\mathbf{x}'\mathbf{e}_j) = \mathbf{a}_k' \underbrace{\boldsymbol{\Sigma}\mathbf{e}_j}_{\lambda_j\mathbf{e}_j} = \mathbf{a}_k'\lambda_j\mathbf{e}_j = \lambda_j e_{jk}$$

CP per le variabili standardizzate

- Le CP si possono ottenere dalle variabili standardizzate $Z=(X-\mu)/\sigma$
- In questo caso si estraggono autovalori e autovettori dalla matrice di correlazione $\boldsymbol{\rho}$
- Autovalori e autovettori di $\boldsymbol{\rho}$ sono diversi da quelli di $\boldsymbol{\Sigma}$
- La varianza totale delle Z è $tr(\boldsymbol{\rho})=p$

Come si ottengono le variabili standardizzate

$$\mathbf{Z}=(\mathbf{V}^{1/2})^{-1}(\mathbf{X}-\boldsymbol{\mu})$$

- $(\mathbf{V}^{1/2})^{-1}=diag(1/\sigma_j)$
- $E(\mathbf{Z})=\mathbf{0}$
- $Cov(\mathbf{Z})=E(\mathbf{Z}\mathbf{Z}')=$
  $=(\mathbf{V}^{1/2})^{-1}(\mathbf{X}-\boldsymbol{\mu})\,(\mathbf{X}-\boldsymbol{\mu})'\,(\mathbf{V}^{1/2})^{-1}=(\mathbf{V}^{1/2})^{-1}\boldsymbol{\Sigma}(\mathbf{V}^{1/2})^{-}=\boldsymbol{\rho}$

CP dalla matrice di correlazione

- Estrazione autovalori e autovettori da $\boldsymbol{\rho}$ $(\lambda_j, \mathbf{e}_j)$, $j=1,2,\ldots,p$
- $Y_j=\mathbf{e}_j'\mathbf{Z}=\mathbf{e}_j'(\mathbf{V}^{1/2})^{-1}(\mathbf{X}-\boldsymbol{\mu})$
- $Var_tot(\mathbf{Z})=tr(\boldsymbol{\rho})=p$, Varianza totale di $\mathbf{Y}$
- Proporzione di varianza spiegata da *j*-ma CP λ_j/p

PROPRIETÀ

$Y_k = e_{k1}Z_1 + e_{k2}Z_2 + \ldots + e_{kp}Z_p = \mathbf{e}'_k\mathbf{z}$

(1) $\mathrm{Var}(Y_k) = \lambda_k$

(2) $\mathrm{Corr}(Y_k, Z_j) = e_{j\sqrt{}}\lambda_1$ $\quad\quad \mathrm{Var}(Z_j) = 1$

(3) $\Sigma\ \lambda_j = \Sigma\ 1 = p \quad j = 1, \ldots, p$

(4) λ e **e** sono gli autovalori e autovettori della matrice di correlazione: $\rho\mathbf{e}_k = \lambda\mathbf{e}_k$

(5) autovettori ortogonali: $\mathbf{E}'\mathbf{E} = \mathbf{I}$, $\mathbf{e}_j'\mathbf{e}_k = 0$

Un semplice esempio

- Confrontare autovalori e autovettori per $\boldsymbol{\Sigma}$ e $\boldsymbol{\rho}$, con

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 4 \\ 4 & 100 \end{bmatrix}$$

le CP estratte da Σ sono diverse dalle CP estratte da ρ

esempio1.sas

CP da Σ o da ρ?

- Nonostante che la relazione tra variabili originalie standardizzate sia semplice, non è possibile ottenere le CP da ρ da quelle da Σ
  - autovalori e autovettori delle due matrici non hanno una relazione semplice!
  - In generale, se si esprimono le CP da ρ facendo la trasformazione inversa $x = \mu + \sigma z$, queste CP NON sono le stesse ricavate da Σ

Perché le CP sono diverse?

- Le CP sono invarianti rispetto a trasformazioni ortogonali di x ma non, in generale, rispetto ad altre trasformazioni
- La transformazione da x a z non è una trasformazione ortogonale
- Inoltre, le matrici di correlazione e covarianza non danno la stessa informazione e non possono essere derivate direttamente una dall'altra.

Quale matrice per definire le CP?

- L'argomento principale a favore della **matrice di correlazione** è che I risultati dell'analisi per insiemi di v.c. diverse sono più direttamente confrontabili
- Il problema maggiorre dell'analisi in CP basata sulla matrice di covarianza è la sensibilità dell CP rispetto alle unità di misura degli elementi di **x**:
  - Se ci sono grosse differenze tra le varianze delle x, allora le variabili con varianza maggiore tendono a dominare le prime CP

Effetto unità di misura

- Sia x_1 una lunghezza, misurabile in cm o in mm.
- Mentre x_2 sia un peso misurato in gr.
- Le matrici di covariaza nei due casi sono

$$\Sigma_1 = \begin{pmatrix} 80 & 44 \\ 44 & 80 \end{pmatrix}, \Sigma_2 = \begin{pmatrix} 8000 & 440 \\ 440 & 80 \end{pmatrix}$$

- La 1ª CP è $Y_{11}=0.707x_1 + 0.707x_2$ per Σ_1 e $Y_{12}=0.998x_1 + 0.055x_2$ per Σ_2
- Cambiando l'unità di misura di x_1 la 1ª CP cambia completamente:
  - Y_{11} assegna lo stesso peso alle due variabili, mentre Y_{12} è dominata da x_1.
  - la 1ª CP 'spiega' il 77.5% della varianza totale di Σ_1, e il 99.3% di Σ_2.

Quando derivare PC da Σ?

- L'inferenza statistica per le CP della popolazione è più semplice per la matrice di covarianza (in pratica però ACP è utilizzata soprattutto a fini descrittivi)
- Quando tutte le x hanno la stessa unità di misura, standardizzare le x equivale a compiere una scelta arbitraria dell'unità di misura.

ANALISI IN COMPONENTI PRINCIPALI dati campionari

- Descrizione dei dati contenuti in una tabella individui-variabili numeriche;
- riduzione delle variabili e rappresentazione grafica: proiezione ortogonale dei punti dallo spazio originale *p*-dimensionale ad uno spazio a *k* dimensioni ($k<p$)

Vedi anche acp_dettagli 10-18

Obiettivo ACP

Ridurre la dimensionalità del problema originale, conservando il massimo di informazione:

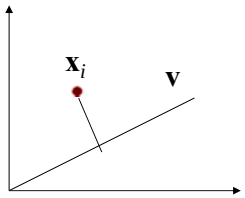
$$\text{VAR TOT}=\Sigma^{p}_{k=1}\sigma_k^2,\ \text{Var}(X_k)=\sigma_k^2$$

- **Criterio**

Scelta del piano di proiezione in modo da conservare al meglio le distanze originali.

Proiezione ortogonale

$\mathbf{x}_i'=(x_{1i},x_{2i})\quad i=1,\ldots,n$



nuova variabile

c_i: $(\mathbf{x}_i - c_i\,\mathbf{v})'\mathbf{v}=0$

$c_i = \mathbf{x}_i'\mathbf{v}$

$(\mathbf{x}_i - c_i\,\mathbf{v}) \perp \mathbf{v}$

- Proiettando i punti $\mathbf{x}_i$ sul vettore $\mathbf{v}$ troviamo: c_i determinazioni della nuova variabile

- *Vediamo* i dati dal punto di vista definito dall'asse $\mathbf{v}$

La prima componente principale

Proiezione ortogonale dei punti $\mathbf{x}_i$ su $\boldsymbol{v}$:

$\mathbf{c}_1 = v_1\mathbf{x}_1 + v_2\mathbf{x}_2 + \ldots + v_p\mathbf{x}_p = \mathbf{Xv}$

combinazione lineare delle $\boldsymbol{X}$ originali

Notare che, posto $M(\mathbf{x}_{(j)})=0$ si ha:

$$\text{Var}(\mathbf{c}) = \frac{1}{n}\mathbf{c}'\mathbf{c} = \frac{1}{n}\mathbf{v}'\mathbf{X}'\mathbf{X}\mathbf{v} = \mathbf{v}'\mathbf{S}\mathbf{v}$$

$\mathbf{x}_i \in R^p,\ \mathbf{x}_{(j)} \in R^n$

Il primo asse principale $\mathbf{c}_1=\mathbf{Xv}_1$

$\mathbf{v}_1$: autovettore associato a λ_1 (autovalore più grande di $\mathbf{S}$)

$\mathbf{X}$ matrice dei dati con variabili centrate

NB: ricordando che $\mathbf{Sv}_1=\lambda_1\mathbf{v}_1$ e $\mathbf{v}'_1\mathbf{v}_1=1$

$\text{Var}(\mathbf{c}_1)=\lambda_1$

$\text{Var}(\mathbf{c}_1)=(1/n)\,\mathbf{c}'_1\,\mathbf{c}_1=(1/n)\,\mathbf{v}'_1\mathbf{X}'\,\mathbf{X}\mathbf{v}_1=\mathbf{v}'_1\underbrace{\mathbf{S}\mathbf{v}_1}_{=\lambda_1\mathbf{v}_1}=$

$=\mathbf{v}'_1\,\lambda_1\,\mathbf{v}_1=\lambda_1\,\underbrace{\mathbf{v}'_1\mathbf{v}_1}_{=1}=\lambda_1$

PROPRIETÀ

(1) $\text{Var}(\mathbf{c}_1)=\lambda_1$

(2) $\text{Corr}(\mathbf{c}_1, \mathbf{x}_j)=[v_j*\text{sqrt}(\lambda_1)]/s_j$, $s_j^2=\text{Var}(\mathbf{x}_j)$

(3) $\Sigma\ \lambda_j=\Sigma\ \sigma^2_j$, $j=1, \ldots, p$

(4) λ e $\boldsymbol{v}$ sono gli autovalori e autovettori della matrice di covarianza: $\mathbf{S}\boldsymbol{v}=\lambda\boldsymbol{v}$

(5) autovettori ortogonali: $\mathbf{V'V}=\mathbf{I}$, $\mathbf{v}_j'\mathbf{v}_k=\mathbf{0}$

(6) Se si utilizza la matrice di correlazione

➔ ACP su variabili standardizzate

Un semplice esempio (Zani p. 223)

- Laureati per voto di maturità e voto di laurea (laureati.txt)
- Calcolare medie e deviazioni standard delle due variabili
- Calcolare matrice di covarianza
- Calcolare autovalori e autovettori
- Verificare che la somma degli autovalori è pari alla varianza totale
- Qual è la proporzione di varianza spiegata dalla prima componente principale?

… segue esempio Zani

- Qual è il peso attribuito alle variabili nella 1ª CP?
- Come possiamo interpretare la 1ª CP?
- Calcolare i punteggi di ciascuno studente lungo la 1ª CP: $y_{1i}=v_{11}(x_{1i}-x_{1.})+v_{12}(x_{2i}-x_{2.})$
- Quale studente presenta la migliore riuscita negli studi?
- Calcolare i coefficienti di correlazione tra la 1ª CP e le due variabili originali $r(Y_1, X_j)=\dfrac{v_{1j}\sqrt{\lambda_1}}{\sqrt{\text{var}(X_j)}}$

Componenti principali nel caso bivariato

(X_1,X_2) bivariata con matrice di correlazione $\mathbf{R}$

$$\mathbf{R}=\begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}$$

Per trovare le componenti principali, dobbiamo trovare autovalori e autovettori di $\mathbf{R}$.

- Gli autovalori sono le radici dell'equazione $|\mathbf{R}-\lambda\mathbf{I}|=0$
- polinomio di 2° in λ: $(1-\lambda)^2-\rho^2=0$, che ammette soluzioni $\lambda_1=1+\rho$ e $\lambda_2=1-\rho$, con $\lambda_1+\lambda_2=2=\text{tr}(\mathbf{R})$
- Autovettore corrispondente a λ_1 si ottiene da:

$$\mathbf{R}\mathbf{v}_1=\lambda_1\mathbf{v}_1$$

- ammette infinite soluzioni per $v_{11}=v_{12}$
- Ponendo il vincolo $\mathbf{v}'_1\mathbf{v}_1=1$, si ottiene $v_{11}=v_{12}=(1/2)^{0.5}$
- In maniera analoga troviamo l'autovettore associato alla seconda radice: $v_{21}=(1/2)^{0.5}$ $v_{22}=-(1/2)^{0.5}$

Componenti principali nel caso bivariato

- Le due componenti principali sono
  $c_1=(1/2)^{0.5}*(x_1+x_2)$, $c_2=(1/2)^{0.5}*(x_1-x_2)$
- Se $\rho<0$, l'ordine degli autovalori e quindi degli autovettori è rovesciato;
- Se $\rho=0$, $\lambda_1=\lambda_2=1$ e qualunque soluzione che formi un angolo retto può essere scelta a rappresentare le due componenti;
- La scelta del segno degli elementi di $\mathbf{v}_j$ è arbitraria: di solito si sceglie $v_{j1}>0$;
- c_1 e c_2 non dipendono dal valore di ρ, tuttavia la proporzione di varianza spiegata cambia con ρ:
  $\tau_1 = \lambda_1 /2=(1+\rho)/2$, $\tau_1 \rightarrow 1$ per $\rho \rightarrow 1$

Esercizio

- Per 25 famiglie con due figli in età adulta è stata misurata la lunghezza della testa dei due figli. Il vettore delle medie e la matrice di covarianza sono:

$$\bar{\mathbf{x}} = \begin{pmatrix} 185.72 \\ 183.84 \end{pmatrix}, \quad S = \begin{pmatrix} 95.29 & \\ 69.66 & 100.81 \end{pmatrix}$$

- Trovare le componenti principali estratte dalla matrice di covarianza.
- Dato il valore delle variabili misurate sulla i-ma famiglia (191, 179), calcolare il valore delle componenti principali corrispondenti.
- Qual è la proporzione di varianza spiegata dalla prima componente principale?

Sintesi della variabilità campionaria attraverso le CP

- La variabilità campionaria può essere sintetizzata attraverso le prime 1,2 o 3 CP?
- Vediamo alcuni esempi
- Esempio 1 JW p. 443

Scelta del numero di componenti

- Tecniche 'formali'
  (Joliffe, 1986; Rencher, 1995)
- Tecniche 'informali':
  - N. di componenti tale da spiegare una proporzione sufficientemente alta della varianza totale (tipo 70% o 90%);
  - Escludere le componenti il cui autovalore è inferiore alla media: $\Sigma\lambda i/p$. Ricordando che $\Sigma\lambda i$=traccia($\mathbf{S}$), $M(\lambda)=M(V(\mathbf{x}_j))$;
  - Scree-plot

Esempio: soddisfazione sul lavoro

	SJ	SJT	SWC	SMC	SDC
SJ	1.00	0.45	0.55	0.20	0.16
SJT	0.45	1.00	0.45	0.25	0.24
SWC	0.55	0.45	1.00	0.30	0.23
SMC	0.20	0.25	0.30	1.00	0.62
SDC	0.16	0.24	0.23	0.62	1.00

Matrice di correlazione

N=9147, p=5
SJ Satisfaction with job
SJT Satisfaction with job training
SWC Satisfaction with working conditions
SMC Satisfaction with Medical care
SDC Satisfaction with Dental Care

princomp1.sas

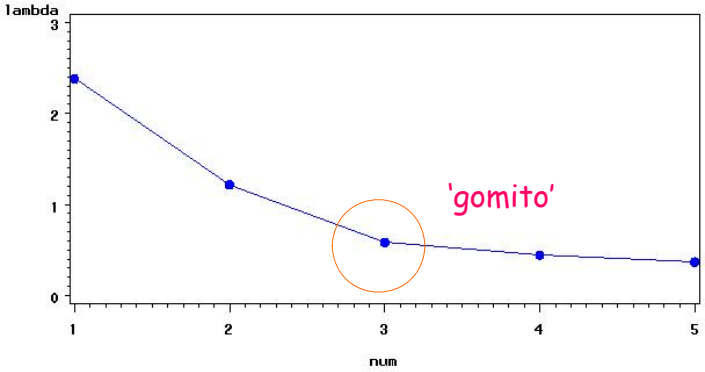
Scelta del n. di componenti: proporzione di varianza spiegata >70%

Autovalori ≥M(λ)=1

```
The PRINCOMP Procedure

Observations        9147
Variables              5

          Eigenvalues of the Correlation Matrix

   Eigenvalue    Difference    Proportion    Cumulative

   2.38266335    1.16695826        0.4765        0.4765
   1.21570509    0.63096612        0.2431        0.7197
   0.58473897    0.13810834        0.1169        0.8366
   0.44663062    0.07636865        0.0893        0.9259
   0.37026197                      0.0741        1.0000
```


Scelta del n. di componenti: scree-plot

Analisi in componenti principali: scree–plot



PROVATE VOI …

- Replicare con SAS, STATA o R l'esempio 8.4 p.445 JW

I punteggi delle componenti principali

- Ogni CP si ottiene come combinazione lineare delle *p* variabili originali
  - $\mathbf{y}_k = \mathbf{e}'_k\mathbf{X}$, k=1,2,…,p
  - di solito si considerano le **x** centrate per ottenere **y** con media 0 ➔ $\mathbf{y}_k = \mathbf{e}'_k[\mathbf{X}-\mathbf{1}M(\mathbf{X})]$
  - $\mathbf{y}_k = \mathbf{e}'_k\mathbf{Z}$, se si considera **R**

Punteggi standardizzati

- Si ottengono dividendo i punteggi per l'autovalore corrispondente alla CP considerata $\mathbf{y}_k = \mathbf{e}'_k\mathbf{X}/\sqrt{\lambda_k}$ di solito
- Consentono di confrontare i coefficienti di componenti principali diverse

Tassi di delinquenza in 16 città americane

Obs	CITTA	OMICIDI	STUPRI	RAPINE	AGGRESS	FURTI	TRUFFE	F_AUTO
1	Atlanta	16.5000	24.8000	106	147	1112	905	494
2	Boston	4.2000	13.3000	122	90	982	669	954
3	Chicago	11.6000	24.7000	340	242	808	609	645
4	Dallas	18.1000	34.2000	184	293	1668	901	602
5	Denver	6.9000	41.5000	173	191	1534	1368	780
6	Detroit	13.0000	35.7000	477	220	1566	1183	788
7	Hartford	2.5000	8.8000	68	103	1017	724	468
8	Honolulu	3.6000	12.7000	42	28	1457	1102	637
9	Houston	16.8000	16.6000	289	186	1509	787	697
10	Kansas City	10.8000	43.2000	255	226	1494	955	765
11	Los Angeles	9.7000	51.8000	186	355	1902	1386	862
12	New Orleans	10.3000	39.7000	266	283	1056	1036	776
13	New York	9.4000	19.4000	522	267	1674	1392	848
14	Portland	5.0000	23.0000	157	144	1530	1281	488
15	Tucson	5.1000	22.9000	85	148	1206	756	483
16	Washington	12.5000	27.6000	524	217	1496	1003	739

princomp6.sas

Tassi di delinquenza in 16 città americane

Variabile	Media	Varianza
OMICIDI	9.75	24.10
STUPRI	27.49	153.69
RAPINE	237.25	24776.87
AGGRESS	196.25	7131.67
FURTI	1375.69	93055.56
TRUFFE	1003.56	68427.60
F_AUTO	689.13	22755.85

n=16 *p*=7

$\mathbf{X}_j$: rapporti su Pop*10.000

Matrice di correlazione

	OMICIDI	STUPRI	RAPINE	AGGRESS	FURTI	TRUFFE	F_AUTO
OMICIDI	1.00	0.35	0.44	0.56	0.23	-.07	0.05
STUPRI	0.35	1.00	0.21	0.76	0.46	0.49	0.36
RAPINE	0.44	0.21	1.00	0.53	0.27	0.26	0.46
AGGRESS	0.56	0.76	0.53	1.00	0.42	0.34	0.38
FURTI	0.23	0.46	0.27	0.42	1.00	0.76	0.27
TRUFFE	-.07	0.49	0.26	0.34	0.76	1.00	0.31
F_AUTO	0.05	0.36	0.46	0.38	0.27	0.31	1.00

Autovalori

	Eigenvalue	Difference	Proportion	Cumulative
1	149233.212	118361.856	0.6899	0.6899
2	30871.356	11638.466	0.1427	0.8326
3	19232.890	6553.719	0.0889	0.9215
4	12679.171	8425.200	0.0586	0.9801
5	4253.971	4209.755	0.0197	0.9997
6	44.216	33.703	0.0002	1.0000
7	10.514		0.0000	1.0000

$\tau_1 = \lambda_1 / \Sigma\lambda_j =$ 0.6899

$\tau_2 = \lambda_1 + \lambda_2 / \Sigma\lambda_j =$ 0.8326

Autovettori

	Prin1	Prin2	Prin3	Prin4	Prin5
OMICIDI	0.001923	0.007882	0.018453	-.011056	0.024173
STUPRI	0.016972	0.011426	-.004771	0.010715	0.120520
RAPINE	0.148857	0.719694	0.172605	-.604712	-.252157
AGGRESS	0.102543	0.207833	0.110994	-.098257	0.953677
FURTI	0.751649	-.226185	0.600418	0.124349	-.088557
TRUFFE	0.615989	-.083458	-.762981	-.175161	0.021307
F_AUTO	0.150386	0.616870	-.121992	0.760453	-.059250

Correlazione tra $\mathbf{y}_1$ e $\mathbf{x}_j$:

$$\frac{e_{1j}\sqrt{\lambda_1}}{\sqrt{s_{jj}}}$$

0.15134 0.52887 0.36532 0.46908 0.95187 0.90968 0.38512

ACP su dati standardizzati ➔autovalori da matrice **R**

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	3.30686705	1.95219619	0.4724	0.4724
2	1.35467086	0.40259088	0.1935	0.6659
3	0.95207998	0.23502487	0.1360	0.8019
4	0.71705512	0.34035578	0.1024	0.9044
5	0.37669934	0.19759961	0.0538	0.9582
6	0.17909973	0.06557180	0.0256	0.9838
7	0.11352792		0.0162	1.0000

princomp8.sas

Interpretazione: autovettori e factor loadings

	Autovettori		Factor loadings	
	Prin1	Prin2	ρ(Prin1,X)	ρ(Prin2,X)
OMICIDI	0.280018	0.600972	0.50921	0.69947
STUPRI	0.434151	-.035297	0.78949	-0.04108
RAPINE	0.348640	0.312334	0.63399	0.36353
AGGRESS	0.468149	0.255861	0.85132	0.29780
FURTI	0.397678	-.383051	0.72317	-0.44583
TRUFFE	0.366098	-.568678	0.66574	-0.66189
F_AUTO	0.316565	-.066683	0.57567	-0.07761

Prin1: media ponderata di tutte le variabili
→ intensità delinquenza

Prin2: reati contro la persona vs reati contro il patrimonio

Interpretazione delle CP campionarie

- CP campionarie centrate $\mathbf{y}_k=[\mathbf{X}-\mathbf{1}\mathrm{M}(\mathbf{X})]\mathbf{e}_k$
- Se $X\sim N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$ allora le CP campionarie sono realizzazioni delle CP della pop con $\mathbf{Y}\sim N_p(\mathbf{0},\boldsymbol{\Lambda})$
- Possono essere interpretate come una traslazione dell'origine nel punto che ha come coordinate le medie campionarie seguita da una rotazione degli assi lungo le direzioni di massima variabilità

acp_dettagli.pdf

Rappresentazione dei punti

- Scatterplot delle prime due CP per
  - verificare assunzione di normalità
  - valutare le 'distanze' tra i soggetti
- Scatterplot delle ultime due CP per identificare outliers

princomp7.sas: standardizzazione variabili e scatterplot

Punteggi delle CP (scores)

CITTA	Prin1	Prin2
Atlanta	-1.16425	1.05591
Boston	-2.08350	-0.08496
Chicago	-0.89599	2.16701
Dallas	1.18419	1.08490
Denver	1.06368	-1.56337
Detroit	1.84215	0.24937
Hartford	-3.28288	-0.29668
Honolulu	-2.09963	-1.90093
Houston	-0.03438	1.26554
Kansas City	1.05966	0.13263
Los Angeles	3.19904	-1.26475
New Orleans	0.81437	0.64496
New York	1.98542	-0.52969
Portland	-0.72831	-1.59418
Tucson	-2.03103	-0.16202
Washington	1.17144	0.7962

$$\hat{\mathbf{y}}_1 = \mathbf{Z}\hat{\mathbf{e}}_1,\ \mathbf{Z}_{n\times p}$$

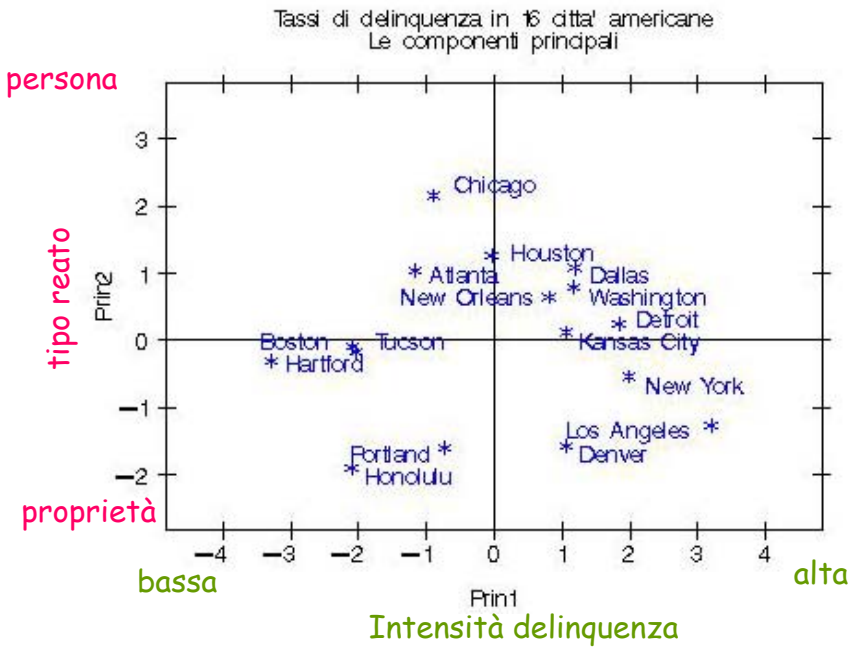
$$\hat{y}_{1i} = e_{11}z_{1i} + e_{12}z_{2i} + \ldots + e_{1p}z_{1p}$$

- $|e_{kj}|$ peso di $\mathbf{x}_j$ in $\mathbf{y}_k$
- segno e_{kj} relazione diretta/inversa tra $\mathbf{x}_j$ e $\mathbf{e}_k$

$$\bar{\hat{\mathbf{y}}}_k = \mathrm{M}(\mathbf{Z}\hat{\mathbf{e}}_1) = \hat{\mathbf{e}}_1'\bar{\mathbf{z}} = 0$$

$$\mathrm{V}(\hat{\mathbf{y}}_k) = \frac{1}{n}\hat{\mathbf{y}}_k'\hat{\mathbf{y}}_k = \hat{\mathbf{e}}_k'\frac{1}{n}\mathbf{Z}'\mathbf{Z}\hat{\mathbf{e}}_k$$

$$= \hat{\mathbf{e}}_k'\mathbf{R}\hat{\mathbf{e}}_k = \hat{\mathbf{e}}_k'\lambda_k\hat{\mathbf{e}}_k = \lambda_k$$



n. di componenti e qualità della rappresentazione

- Per ciascuna variabile, la proporzione di varianza spiegata dalle componenti principali si ottiene elevando al quadrato i coefficienti di correlazione $r_{kj}=\frac{v_{1j}\sqrt{\lambda_1}}{\sigma_j}$ tra la componenti e la variabile
- Per valutare quanto si perde rinunciando a estrarre delle componenti si può calcolare la proporzione di varianza spiegata che e' uguale alla somma degli indici di determinazione lineare r^2_{kj} tra la componente *k*-esima e la variabile *j*-esima.

Quanto si perde rinunciando ad alcune CP? Calcolo di R^2 parziale

NAME	R2CP1	R2CP2	R2CP3	TOT2	TOT3
OMICIDI	0.2593	0.4893	0.0962	0.74856	0.84479
STUPRI	0.6233	0.0017	0.0857	0.62499	0.71065
RAPINE	0.4019	0.1322	0.2208	0.53410	0.75493
AGGRESS	0.7247	0.0887	0.0192	0.81343	0.83260
FURTI	0.5230	0.1988	0.0477	0.72174	0.76946
TRUFFE	0.4432	0.4381	0.0023	0.88130	0.88361
F_AUTO	0.3314	0.0060	0.4802	0.33742	0.81758

princomp8.sas

Sintesi di indicatori con la 1ª C.P

Indicatori di Tagliacarne per la stima dei consumi provinciali

```
X1='SPESA P-C RAI'
X2='SPESA P-C TABACCHI'
X3='SPESA P-C SPETTACOLI'
X4='ABBONATI TEL *1000 AB'
X5='INDICE DI LETTURA'
X6='TRASPORTI PRIV*1000AB';
```

N=20 regioni italiane, p=6 indicatori

princomp5.sas

Correlazione e Autovalori

	x1	x2	x3	x4	x5	x6
x1	1.0000	0.7709	0.7693	0.7583	0.8560	0.8424
x2	0.7709	1.0000	0.7803	0.7282	0.8712	0.7778
x3	0.7693	0.7803	1.0000	0.7857	0.8831	0.8771
x4	0.7583	0.7282	0.7857	1.0000	0.9023	0.7298
x5	0.8560	0.8712	0.8831	0.9023	1.0000	0.8433
x6	0.8424	0.7778	0.8771	0.7298	0.8433	1.0000

Eigenvalues of the Correlation Matrix: Total = 6 Average = 1

	Eigenvalue	Difference	Proportion	Cumulative
1	5.06310584	4.75112135	0.8439	0.8439
2	0.31198449	0.05618323	0.0520	0.8958
3	0.25580126	0.02297661	0.0426	0.9385
4	0.23282465	0.13973662	0.0388	0.9773
5	0.09308803	0.04989230	0.0155	0.9928
6	0.04319573		0.0072	1.0000

6 factors will be retained by the NFACTOR criterion.

Autovettori, corr(**c**,**x**) e r^2 parziali

		Eigenvectors	Factor Pattern
		1	Factor1
x1	SPESA P-C RAI	0.40295	0.90670
x2	SPESA P-C TABACCHI	0.39732	0.89402
x3	SPESA P-C SPETTACOLI	0.41133	0.92556
x4	ABBONATI TEL *1000 AB	0.39543	0.88978
x5	INDICE DI LETTURA	0.43222	0.97254
x6	TRASPORTI PRIV*1000AB	0.40913	0.92060

Final Communality Estimates: Total = 5.063106

x1	x2	x3	x4	x5	x6
0.82210449	0.79928036	0.85665968	0.79170796	0.94584104	0.84751232

Graduatoria delle regioni in base alla prima CP

regione	Factor1
CALABRIA	-1.57671
BASILICATA	-1.47325
CAMPANIA	-1.24921
MOLISE	-1.01093
PUGLIA	-0.99909
SICILIA	-0.95596
SARDEGNA	-0.73558
ABRUZZI	-0.52827
VENETO	-0.06595
MARCHE	0.15236
UMBRIA	0.17913
TRENTINO_A_A	0.32710
LAZIO	0.56126
LOMBARDIA	0.61907
PIEMONTE	0.82196
FRIULI_V_G	0.90359
TOSCANA	0.91966
.....	
VALLE_AOSTA	1.61943

$$\hat{\mathbf{y}}_1 = \mathbf{Z}\hat{\mathbf{e}}_1, \mathbf{Z}_{n\times p}$$

$$\hat{y}_{1i} = e_{11}z_{1i} + e_{12}z_{2i} + \ldots + e_{1p}z_{1p}$$

ACP con STATA

pca [varlist] [weight] [if exp] [in range] [, factors(#) mineigen(#) covariance means]

score newvarlist [if exp] [in range] greigen [, plot(plot) connected_options twoway_options]

- score crea nuove variabili che contengono le cp
- greigen produce il grafico degli autovalori (scree-plot)
- factors(#) indica il n. max di componenti da considerare
- mineigen(#) valore min degli autovalori da considerare
- covariance richiede ACP sulla matrice di covarianza

pca.do

ACP in STATA (2)

- pca non può essere utilizzata direttamente sulla matrice di covarianza o di correlazione.

Come si procede in questo caso?

- Generare variabili corrispondenti alla matrice di covarianza o di correlazione
- utilizzare corr2data e quindi pca.

pca.do

Bibliografia

- Dunteman G. H., 1989, *Principal Component Analysis*, Sage.
- Marchetti G., 1997, *Introduzione all'analisi statistica dei dati multivariati*, dispense ad uso degli studenti, Dipartimento di Statistica "G. parenti", Firenze
- Mardia K.V., Kent J. T., e Bibby J. M., 1979, *Multivariate analysis*, London: Academic Press
- Jolliffe, I.T. , 2002, *Principal Component Analysis*, Springer Series in Statistics
- Bartholomew, D.J., Steele, F., Moustaki, I., Galbraith, J. (2002) *The analysis and interpretation of multivariate data for social sciences*. Chapman and Hall/CRC.

Alcuni articoli recenti che usano ACP

- Pentecost E. J., Van Hooydonk C., Van Poeck A., 2001, Measuring and estimating exchange market pressure in the EU, *Journal of the International Money and Finance*, **20**, 401-418.
- Antonio Minguzzi and Renato Passaro, 2001, The network of relationships between the economic environment and the entrepreneurial culture in small firms**,** *Journal of Business Venturing,* **16**, Pages 181-207.
- Leonidas C. Leonidou, 2000, Barriers to export management; an organizational and internationalization analysis**,** *Journal of International Management,* **6***,* 121-148.
- Robert E. Morgan and Constantine S. Katsikeas, 1998, Exporting Problems of Industrial Manufacturers**,** *Industrial Marketing Management,* **27**, 161-176.
- H. L. Brossard**,** 1998, Information Sources Used by an Organization During a Complex Decision Process; An Exploratory Study**,** *Industrial Marketing Management,* **27**, 41-50

Analisi in componenti principali

- L'ACP può essere utilizzata per
  - Riduzione dei dati (proiezione in dimensione q<p)
  - Trasformare un insieme di variabili correlate in un insieme di variabili incorrelate
  - Trovare combinazioni lineari dei dati con varianza alta (o bassa)

Gara di heptathlon

- Descrizione dei dati in heptathlon.pdf
- Dati in heptathlon.xls
  - Importare i dati in SAS
  - Fare un po' di analisi descrittive (medie, varianze e matrice di correlazione)
  - Estrarre le componenti principali (non usare la variabile score): da S o da R?
  - Commentare i risultati

Esempio dati HATCO
(Hair et al., 1998, pp. 120-134)

- Dati provenienti da studio di segmentazione clienti della HATCO
- 100 osservazioni
- 14 variabili relative alla soddisfazione e alle caratteristiche dei clienti (aziende)

`hatco.sas` crea SAS Data Set

Obiettivo analisi

- chi sono i clienti ;
- Cosa i clienti pensano della HATCO

Utilizzo variabili x1-x4, x6, x7(percezione dei clienti di HATCO)

Riduzione dei dati per avere un'*immagine* della situazione (ACP)

Individuazione fattori latenti (FA)

Analisi preliminare

- Correlazioni: 7 delle 15 correlazioni sono significative al livello 0.01;
- Correlazioni parziali e misura di Kaiser (Measure of Sampling Adequacy, MSA).

0<=MSA<=1
MSA=1 se *x* prevista senza errore dalle altre variabili;
MSA>=0.8 ➔ottimo; 0.8<MSA<=0.7 ➔medio
0.7<MSA<=0.6 ➔mediocre; 0.6<MSA<=0.5 ➔sufficiente;
MSA<0.5 ➔inaccettabile

hatco2.sas

Estrazione delle C.P.

Analisi preliminare: quante componenti?

- 2 per criterio radici latenti (>1);
- 3 per scree-plot;
- Proporzione varianza spiegata 71% per 2 CP

`hatco3.sas`

Interpretazione delle componenti

- Factor loadings Cov(X,CP)
- Final communality 2 var(X) spiegata dalle CP
- Eigenvalues Var(CP)
- Proportion % varianza totale spiegata da ogni CP
- Cumulative % varianza totale spiegata dalle 2 CP

Rotazione delle CP

VARIMAX rotazione ortogonale:

$\Gamma^*=\Gamma H'$, $H=(\cos\alpha, -\text{sen}\alpha; \text{sen}\alpha, \cos\alpha)$

NB.

- Redistribuzione varianza componenti;
- Diverso valore delle correlazioni;
- migliore interpretazione delle CP: nessuna var significativa su entrambe le CP, x5 esclusa
  - 1a CP: x1,x3 vs. x2,x7 valore di base
  - 2a CP: x4,x6 immagine

`hatco4.sas`

PROCEDURE SAS PER L'ANALISI MULTIVARIATA

Riduzione di dimensionalità

PRINCOMP	analisi in componenti principali
PRINQUAL	ACP per dati qualitativi e MDP
MDS	Multidimensional scaling

Relazione tra due insiemi di variabili

CANCORR	Analisi delle correlazioni canoniche
CANDISC	analisi discriminante canonica
CORRESP	analisi delle corrispondenze semplici e multiple.

ESEMPI ACP in SAS

Gli esempi si trovano in v:\carla\acp

`Princomp1.sas`	input: matrice di correlazione; autovalori, autovettori e % var tot.
`Princomp2.sas`	correlazione tra C.P. e X.
`Princomp3.sas`	quante C.P. ? Scree Plot con FACTOR e PRINCOMP.
`Princomp4.sas`	input: dati grezzi; grafico; ordinamento obs.
`Princomp5.sas`	sintesi di indicatori mediante la prima C.P.

Segue ESEMPI ACP IN SAS

`princomp6.sas`	esempio criminalità
`princomp7.sas`	standardizzazione var
`princomp8.sas`	quante C.P? R2 parziale
`princomp9.sas`	utilizzo di PROC FACTOR
`princomp10.sas`	calcola scores, grafico
`princomp11.sas`	analisi ponderata